

빅데이터 분석



소프트웨어융합대학원
진혜진

01. 빅데이터 산업의 이해

■ 빅데이터 산업을 설명하는 용어

- 빅데이터 산업은 관련된 여러 분야가 유기적으로 결합된 시스템
- 빅데이터 플랫폼
 - 데이터 관점에서 빅데이터를 수집·저장·분석하는 프로세스와 그에 필요한 자원의 유기적 결합을 나타냄
- 빅데이터 에코시스템
 - 빅데이터 플랫폼에 서비스 산업을 결합하여 고객에게 가치를 전달 하는 유기적 공동체를 나타냄
- 빅데이터 서비스 프레임워크
 - 빅데이터 에코시스템에서 서비스 공급자를 분류하고 서비스 유형과 수준을 파악한 것을 나타냄

01. 빅데이터 산업의 이해

■ 빅데이터 플랫폼

• 데이터 플랫폼의 발전

- 데이터 플랫폼은 정형화된 형태로 데이터를 저장하는 파일 시스템으로 시작
- 이후에 다수가 동시에 사용할 수 있는 데이터베이스와 데이터 웨어하우스(DW)로 발전
- 폭발적으로 증가하는 데이터를 저장 및 유통하기 위한 빅데이터 플랫폼으로 진화

• 빅데이터 플랫폼의 개념

- 빅데이터를 처리하는 것
- 대량의 데이터를 저장 및 분석, 처리할 수 있는 대용량의 고속 저장 공간과 고성능 계산 능력의 컴퓨팅 인프라를 보유
- 실시간으로 발생하는 빅데이터를 처리 및 분석하여 일관성을 유지하는 데이터 분석도 필요
- 빅 데이터에서 발생하는 개인 정보를 위한 정보 보안 관리체계 지원도 필요
- 빅데이터 플랫폼은 오픈 소스인 하둡을 근간으로 많이 사용

01. 빅데이터 산업의 이해

■ 빅데이터 플랫폼

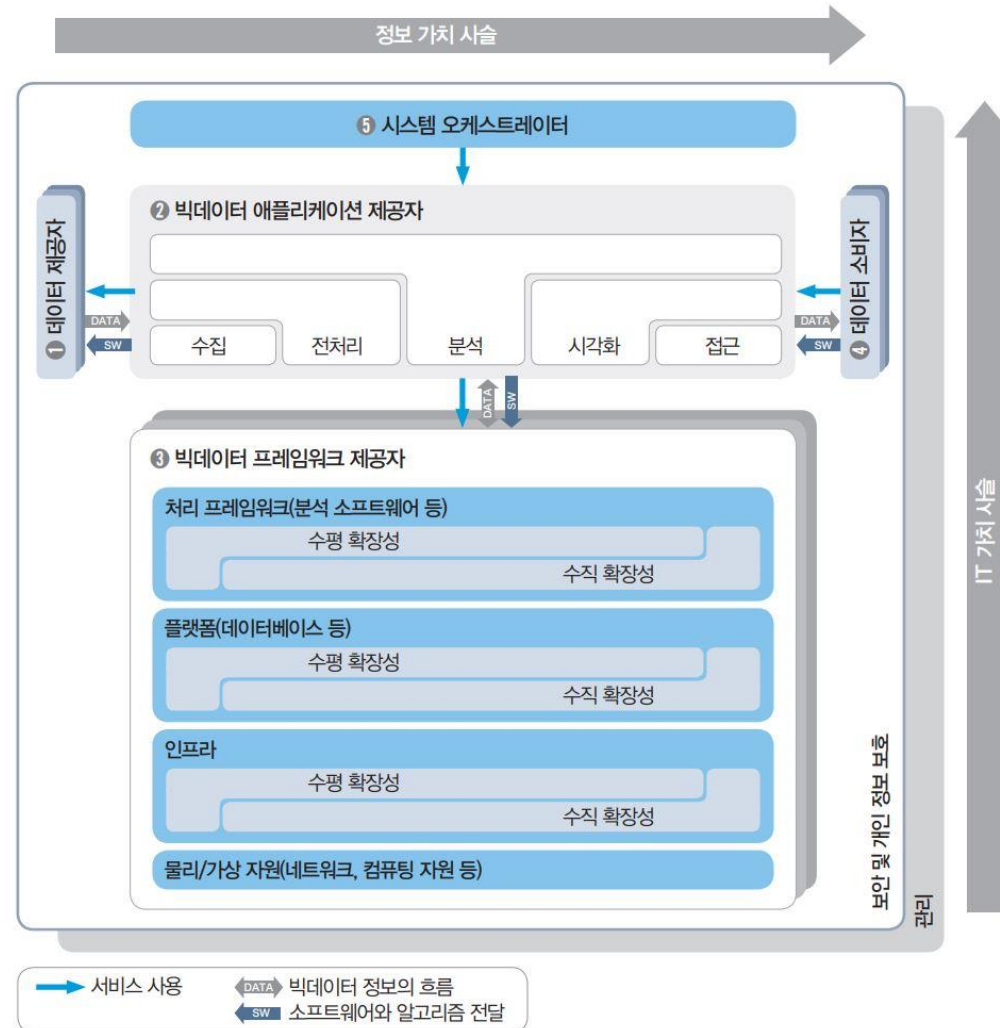


그림 3-1 대표적인 표준화 빅데이터 플랫폼인 NIST의 빅데이터 참조 아키텍처

01. 빅데이터 산업의 이해

■ 빅데이터 서비스 프레임워크

- 빅데이터 서비스 프레임워크는 빅데이터 시장을 효율적으로 이해하기 위한 것
- 에코시스템 안에서 서비스 공급자를 분류하고 서비스 유형과 수준을 파악하는 것이 필요
- 공급하는 서비스의 유형과 수준에 따라 빅데이터 서비스 공급자와 애플리케이션 공급자로 분류

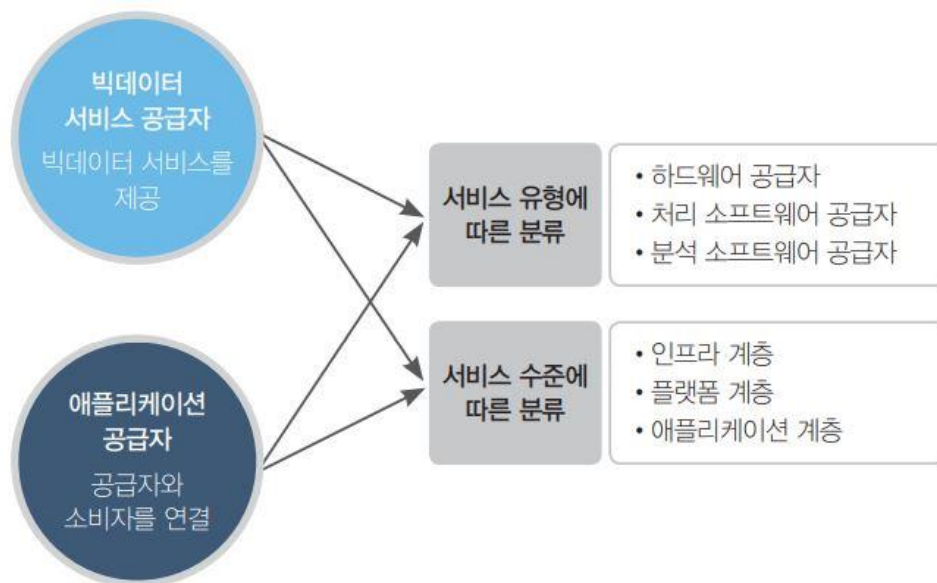


그림 3-3 빅데이터 시장의 공급자 분류

01. 빅데이터 산업의 이해

■ 빅데이터 서비스 프레임워크

■ 공급 서비스 유형에 따른 분류

- 하드웨어 공급자
 - 자체 데이터센터 및 클라우드 시스템을 통해 빅데이터 서비스를 위한 인프라를 공급
- 처리 소프트웨어 공급자
 - 서비스 소비자가 저장한 빅데이터를 효과적으로 저장 및 처리할 수 있는 소프트웨어를 제공한다.
- 분석 소프트웨어 공급자
 - 서비스 소비자의 빅데이터를 분석할 소프트웨어를 제공

■ 공급 서비스 수준에 따른 분류

- 인프라 계층
 - 빅데이터를 위한 기초 작업을 담당하는 하드웨어나 운영체제를 제공
 - 자체 인프라를 구축하거나 가상화를 위한 클라우드 컴퓨팅 서비스가 여기에 속함
- 플랫폼 계층
 - 클라우드 컴퓨팅 서비스나 하드웨어에 종속되지 않는 처리 및 분석 소프트웨어 등을 제공
- 애플리케이션 계층
 - 소비자가 빅데이터와 소통하는 매커니즘을 제공한다. 빅데이터 처리 결과를 바탕으로 소비자가 원하는 분석 결과를 제공하거나 시장에 유통

01. 빅데이터 산업의 이해

■ 빅데이터 서비스 프레임워크

- 빅데이터 서비스 공급자 분류를 위한 빅데이터 서비스 프레임워크

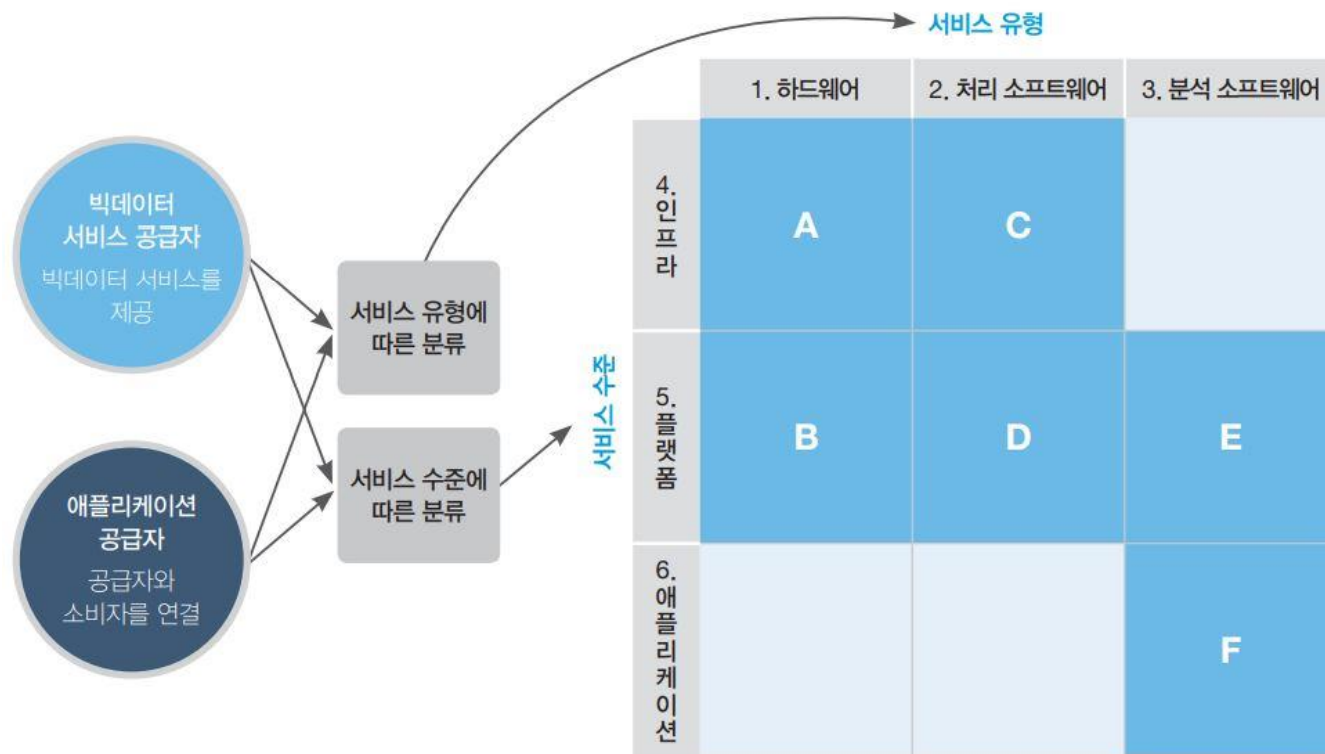


그림 3-4 빅데이터 서비스 공급자 분류를 기반으로 나타낸 빅데이터 서비스 프레임워크

01. 빅데이터 산업의 이해

■ 빅데이터 서비스 프레임워크

- A: 하드웨어-인프라 유형
 - 기업 등에서 자체 데이터센터를 구축할 수 있게 해주는 서비스 유형
 - 이 유형은 사적 데이터를 중심으로 하는 기업형 솔루션과 공적 데이터를 중심으로 하는 플랫폼 서비스로 구분할 수 있음
 - IBM, HP, 오라클 등의 기업용 하드웨어 솔루션 제품이 여기에 해당
- B: 하드웨어-플랫폼 유형
 - 클라우드를 기반으로 서비스를 제공하는 유형
 - 기존의 클라우드 컴퓨팅 시스템을 사용해 빅데이터 서비스를 제공
- C: 처리 소프트웨어-인프라 유형
 - 하드웨어와 소프트웨어를 함께 제공하는 서비스 유형
 - 대용량 데이터를 다루기 위해 필요한 분산 저장 및 병렬 처리 인프라에 처리 솔루션까지 제공
 - 기업용 솔루션 사업을 하는 오라클, IBM, HP, EMC 등의 기업에서 자사의 하드웨어와 특화된 소프트웨어를 통합해서 제공

01. 빅데이터 산업의 이해

■ 빅데이터 서비스 프레임워크

- D: 처리 소프트웨어-플랫폼 유형
 - 오픈 소스 기반의 소프트웨어 플랫폼을 제공하는 서비스 유형
 - 공급자는 오픈 소스를 기반으로 하는 빅데이터 처리 프로그램을 공급
 - 소비자는 공급자가 제공하는 클라우드 서비스를 통해 빅데이터 처리 서비스를 이용할 수 있음
- E: 분석 소프트웨어-플랫폼 유형
 - 일반 소비자를 위한 분석 소프트웨어를 제공하는 서비스 유형
 - 빅데이터를 솔루션으로 상품화하고 클라우드 컴퓨팅과 결합하여 제공
 - 소비자는 자체 서버와 솔루션을 구축하는 대신에 클라우드 컴퓨팅 인프라에서 데이터를 저장 및 분석하는 프로그램을 이용할 수 있음
- F: 분석 소프트웨어-애플리케이션 유형
 - 고객 맞춤형 솔루션 서비스 유형으로 데이터의 의미를 파악하고 이를 분석해서 활용하는 서비스를 제공
 - 축적된 데이터를 바탕으로 분석 후 결과의 의미를 파악하여 제공
 - 소비자의 검색 패턴을 이용해 독감 확산을 예측했던 구글 분석이 대표적 사례

02. 빅데이터 분석 방법과 접근법

■ 빅데이터 분석 방법

■ 분석 목적에 따른 구분

① 통계 분석

- 통계 기법에 의한 분석 방법으로 가장 대표적인 유형

② 예측 분석

- 과거의 데이터와 변수 간의 관계를 이용하여 새로운 변수를 추정

③ 데이터 마이닝 분석

- 많은 데이터 속에 숨겨진 유용한 패턴을 추출하여 분류, 군집, 연관, 이상 탐지 분석 등을 수행

④ 최적화 분석

- 주어진 제한 조건을 만족하면서 목적 함수를 최대화 또는 최소화하는 방법 을 찾는다.

02. 빅데이터 분석 방법과 접근법

■ 빅데이터 분석 접근법

• 하향식 접근법

- 문제 해결 방법을 찾기 위해 필요한 데이터를 수집 및 분석하는 방식
- 문제 해결을 위해 근본 원인을 파악하고 분석 과제를 도출한 뒤 해결 방안을 도출
- 도출된 해결 방안에 대한 실현 가능성과 우선순위를 결정하기 위해 데이터를 수집, 가공, 분석하는 접근법
- 분석 과제를 도출하기 위해 '수요 기반 분석 과제 도출 방식'을 사용
- 데이터 분석은 문제 해결을 가능하게 하는 실행 동인 역할

• 상향식 접근법

- 현재 보유하고 있는 데이터를 분석하여 의미 있는 관계나 패턴을 찾아 지식을 발견하고 문제를 해결하는 방식
- 정형 데이터는 물론이고 다양한 원천의 비정형 데이터를 조합 하고 시각화를 통해 의미 있는 패턴을 파악한 뒤 이를 적용하여 문제를 해결하는 데이터 기반의 접근
- 분석 과제를 도출하기 위해 '데이터 주도 분석 과제 도출 방식'을 사용
- 데이터는 추진 동인 역할

• 프로토타이핑 접근법

- 빅데이터 환경의 불확실성을 고려한 방식
- 소비자의 요구 사항이나 데이터를 규정하기가 어렵고 데이터 원천도 명확히 파악하기 어려운 경우 사용
- 일단 프로토타입을 만들어 분석을 시도한 뒤 결과를 확인하고 개선하고 이를 반복

03. 빅데이터 분석을 위한 데이터 과학 방법론

■ 데이터 과학 방법론

- 여섯 단계로 구성되며 필요에 따라 특정 단계를 반복해서 수행 가능



그림 3-5 데이터 과학 방법론의 6단계 구성

프로젝트 헌장(Project Charter)					
프로젝트 명 (Project Name)					
프로젝트 설명 (Project Description)					
프로젝트 매니저 (Project Manager, PM)		승인 날짜 (Date Approved)			
프로젝트 스폰서 (Project Sponsor)		서명 (Signature)			
비즈니스 케이스(Business Case)		목표(Goals) / 산출물(Deliverables)			
팀 구성원(Team Member)					
이름(Name)	역할(Role)				
위험과 제약사항(Risk and Constraints)		주요 일정(Milestones)			

03. 빅데이터 분석을 위한 데이터 과학 방법론

■ [2단계] 데이터 수집

- 프로젝트에 필요한 데이터의 위치와 형태를 확인하고 원시 데이터를 수집
 - 필요한 데이터를 수집할 때는 이미 가지고 있는 내부 데이터베이스나 데이터 저장소를 이용
 - 외부에서 수집하는 경우 다양한 수집 기술을 활용할 수 있음
 - 수집할 데이터의 유형과 종류를 파악한 뒤 그에 맞는 수집 기술을 선택해서 사용

표 3-2 개방 데이터를 제공하는 사이트

사이트	설명
http://data.go.kr	한국 정부에서 제공하는 공공데이터
http://kostat.go.kr	한국 통계청에서 공개하는 데이터
http://opendata.hira.or.kr	한국 보건 의료 빅데이터 개방 시스템
http://www.localdata.kr	한국 지방행정 인허가 데이터
https://www.mcst.go.kr	한국 문화체육관광부 문화 데이터
http://data.seoul.go.kr	서울시 열린데이터 광장
https://data.gg.go.kr	경기도 공공데이터 개방 포털
http://data.gov	미국 정부의 공공데이터
http://data.worldbank.org	세계 은행에서 제공하는 개방 데이터
http://open.fda.gov	미국 식약청의 개방 데이터

표 3-4 데이터의 유형과 종류에 따라 사용할 수 있는 수집 기술의 예

유형	종류	수집 기술
정형 데이터	RDB, 스프레드시트	ETL, FTP, Open API
반정형 데이터	HTML, XML, JSON, 웹 문서, 웹 로그, 센서 데이터	크롤링, RSS, Open API, FTP
비정형 데이터	소셜 데이터, 문서(워드, 한글), 이미지, 오디오, 비디오, IoT	크롤링, RSS, Open API, 스트리밍, FTP

표 3-3 다양한 데이터 수집 기술

수집 기술	설명	수집 데이터
크롤링	• SNS, 뉴스, 웹 정보처럼 인터넷에서 제공하는 데이터를 수집할 수 있다.	웹 추출 데이터
FTP	• TCP/IP 프로토콜을 활용하는 인터넷 서버에서 각종 파일을 송수신할 수 있다. • 보안을 강화하려면 SFTP 사용을 고려해야 한다. • 서버 간 연동시에는 전용 네트워크 구축을 고려해야 한다.	파일
Open API	• 서비스, 데이터 등을 어디서나 쉽게 이용하도록 개방된 API로 데이터 수집 방식을 제공한다. • 다양한 애플리케이션을 개발할 수 있도록 개발자와 소비자에게 공개되어 있다.	실시간 수집 데이터
RSS	• 웹 기반의 최신 정보를 공유하기 위한 XML 기반의 콘텐츠 배급 프로토콜이다.	XML 기반 웹 콘텐츠
스트리밍	• 인터넷에서 실시간으로 음성/오디오/비디오 데이터를 수집하는 기술이다.	음성/오디오/비디오의 실시간 수집 데이터
로그 수집기	• 웹 서버 로그, 웹 로그, 트랜잭션 로그, 클릭 로그, DB 로그 등 각종 로그 데이터를 수집하는 오픈 소스 기술이다. • Chukwa, Flume, Scribe 등이 있다.	로그
RDB 수집기	• 관계형 데이터베이스에서 정형 데이터를 수집한 뒤 HDFS(하둡 분산 파일 시스템)나 HBase와 같은 NoSQL에 저장하는 오픈 소스 기술이다. • Sqoop, Direct JDBC/ODBC 등이 있다.	RDB 기반 데이터

03. 빅데이터 분석을 위한 데이터 과학 방법론

■ [3단계] 데이터 준비

- 수집한 원시 데이터의 품질을 높이기 위해 정제 후 사용 가능한 형태로 가공하는 단계
- 수집한 데이터를 다음 단계에서 사용할 수 있게 오류를 여과 하거나 수정하여 정제
- 필요에 따라서는 데이터를 통합하거나 형태를 변환

■ [4단계] 데이터 탐색

- 데이터와 변수 간의 관계나 상호 작용을 이해하기 위한 단계
- 변수 간의 관련성, 데이터의 분포, 편차, 패턴 존재 여부를 확인하는 탐색적 데이터 분석(EDA)이라고도 함
- 데이터를 쉽게 이해하기 위해 꺾은선 그래프나 히스토그램, 분포도 등과 같은 그래픽 기법을 많이 사용

표 3-5 데이터 준비에 필요한 작업

종류	설명
데이터 여과	• 오류 발견, 보정, 삭제, 중복성 확인 등의 과정을 통해 데이터 품질을 향상시킨다.
데이터 정제	• 결측치는 채워 넣고 이상치는 식별 또는 제거하고 잡음이 섞인 데이터는 평활화하여 데이터 불일치성을 교정한다.
데이터 통합	• 데이터 분석이 용이하도록 유사 데이터 및 연계가 필요한 데이터(또는 데이터베이스)를 통합한다.
데이터 축소	• 분석 시간을 단축하기 위해 분석에 사용하지 않는 항목은 제거한다.
데이터 변환	• 데이터 분석에 용이한 형태로 데이터 유형을 변환한다. • 정규화normalization, 집합화aggregation, 요약summarization, 계층 생성 등의 방법을 활용한다. • ETLExtraction, Transformation, Loading 도구를 제공한다.

03. 빅데이터 분석을 위한 데이터 과학 방법론

■ [5단계] 데이터 모델링

- 이전 단계에서 얻은 데이터 탐색 결과로 프로젝트에 대한 답을 찾는 단계
- 변수를 선택하여 모델을 구성하고 실행 및 평가하는 과정을 반복 수행하여 문제 해결 모델을 완성
- 이때 분석하려는 데이터의 특성과 목적에 따라 모델 유형을 선택할 수 있음

표 3-6 데이터 분석 모델의 종류

유형	종류 및 설명
통계 분석 모델	전통적인 분석 기법이다. 주로 수치형 데이터에 사용하며 확률을 기반으로 현상을 추정 및 예측한다.
	기술 통계 대표적인 것으로 평균(산술평균, 중앙값, 최빈값), 분산, 표준편차가 있다.
	상관 분석 두 변수가 어떤 선형적 관계를 가지는지 분석하는 기법이다. 두 변수는 서로 독립적 관계일 수도 있고 상관된 관계일 수 있는데 이러한 관계의 강도를 상관관계라고 한다.
	회귀 분석 연속형 변수에 대해 독립 변수와 종속 변수 사이의 상관관계에 따른 수학적 모델인 선형적 관계식을 구하여 어떤 독립 변수가 주어졌을 때 이에 따른 종속 변수를 예측하거나 수학적 모델이 얼마나 잘 설명하고 있는지를 판별하기 위한 적합도를 측정하는 분석 기법이다.
	분산 분석 두 개 이상 다수의 집단을 비교할 때 집단 내의 분산, 총평균과 각 집단의 평균의 차이로 생긴 집단 간 분산의 비교를 통해 만들어진 F분포로 가설을 검증하는 기법이다.
	주성분 분석 다양한 변수를 분석하는 다변량 분석으로 많은 변수로부터 몇 개의 주성분을 추출하는 기법이다. 이때 주성분 분석은 차원 축소를 위한 것이다.
데이터 마이닝 모델	패턴 인식, AI, 머신러닝, 딥러닝 등을 이용하여 대용량 데이터에 숨겨진 데이터 간의 상호 관련성 및 유용한 정보를 추출하는 기법이다.
	예측 대용량 데이터 집합 내의 패턴을 기반으로 미래를 예측한다(예: 수요 예측).
	분류 일정한 집단에 대해 특정한 정의로 분류 및 구분을 추론한다.
	군집화 구체적인 특성을 공유하는 자료를 분류한다. 미리 정의된 특성에 대한 정보를 가지지 않는다는 점에서 분류와 다르다(예: 유사 행동 집단의 구분).
	패턴 분석 동시에 발생한 사건 간의 상호연관성을 탐색한다(예: 장바구니 속 상품의 관계).
	순차 패턴 분석 연관 규칙에 시간 개념을 반영하여 시계열에 따른 패턴의 상호연관성을 탐색한다(예: 금융 상품 사용을 위한 반복 방문).
텍스트 마이닝 모델	텍스트 기반의 데이터로부터 새로운 정보를 발견할 수 있도록 정보 검색, 추출, 체계화, 분석을 모두 포함하는 텍스트 처리 과정 및 기법이다.
소셜 네트워크 분석 모델	언어 분석 기반의 정보 추출을 통해 대용량의 소셜 미디어 데이터에서 이슈를 탐지하고 시간 경과에 따라 이슈가 유통되는 전체 과정을 모니터링하고 향후 추이를 분석하는 기법이다.

03. 빅데이터 분석을 위한 데이터 과학 방법론

■ [6단계] 결과 발표 및 분석 자동화

- 프로젝트 수행 결과가 연구 목표를 달성했는지를 이해 당사자, 특히 의사 결정자에게 이해시키고 가능하다면 이후의 유사 프로젝트 수행을 위해 분석 과정을 자동화하는 단계
- [1단계]에서 작성한 프로젝트 헌장에 명시된 목표를 달성했는지 산출물이 제대로 작성되었는지, 일정과 예산은 계획대로 진행되었는지 여부를 확인
- 모든 참여자를 대상으로 분석 결과를 발표
- 분석 과정을 재사용할 수 있도록 자동화